A Study of Multi-Agent Collaboration Theories

Burt Wilsker

ISI/RR-96-449

November, 1996

USC
UNIVERSITY OF SOUTHERN CALIFORNIA

INFORMATION SCIENCES INSTITUTE

School of Engineering / 4676 Admiralty Way, Suite 1001
Marina del Rey, California 90292-6695 / 310 822 1511

ISI Research Report
ISI/RR-96-449
November, 1996

A Study of Multi-Agent Collaboration Theories

Burt Wilsker

ISI/RR-96-449

November, 1996

Burt Wilsker's current contact information:

Jet Propulsion Laboratory, 4800 Oaks Grove Drive, Pasadena CA 91109
Wilsker@csi.jpl.nasa.gov

University of Southern California
Information Science Institute
4676 Admiralty Way, Marina del Rey, CA. 90292-6695
310-822-1511

REPORT DOCUMENTATION PAGE

FORM APPROVED
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 1996	Research Report

4. TITLE AND SUBTITLE

A Study of Multi-Agent Collaboration Theories

5. FUNDING NUMBERS

None

6. AUTHOR(S)

Burt Wilsker

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

USC INFORMATION SCIENCES INSTITUTE
4676 ADMIRALTY WAY
MARINA DEL REY, CA 90292-6695

8. PERFORMING ORGANIZATON REPORT NUMBER

ISI/RR-96-449

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

None

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12A. DISTRIBUTION/AVAILABILITY STATEMENT

UNCLASSIFIED/UNLIMITED

12B. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This paper analyzes several theories of multi-agent collaboration. We have chosen three to study more closely than the others. While these contain what we believe to be the essential elements of a multi- agent collaboration, each approaches this idea somewhat differently. This is followed by a less-extensive review of several other theories, a number of which allow for contractual relationships. This analysis reveals not only relative strengths and weaknesses between the theories, but also illustrates the differences when agents collaborate and when they have a contractual association, the line between which is often blurred in the literature. Finally, we present two "real-world" domains, and discuss the advantages and disadvantages of these theories to each domain.

14. SUBJECT TERMS

distributed artificial intelligence, multi-agent collaboration,

15. NUMBER OF PAGES

24

16. PRICE CODE

17. SECURITY CLASSIFICTION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reoprts. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to stay within the lines to meet optical scanning requirements.

Block 1. Agency Use Only (Leave blank).

Block 2. Report Date. Full publication date including day, month,a nd year, if available (e.g. 1 jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element numbers(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	PR - Project
G - Grant	TA - Task
PE - Program Element	WU - Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the repor.

Block 9. Sponsoring/Monitoring Agency Names(s) and Address(es). Self-explanatory

Block 10. Sponsoring/Monitoring Agency Report Number. (If known)

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of ...; To be published in... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.
DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - Leave blank.
NTIS - Leave blank.

Block 13. Abstract. Include a brief (Maximum 200 words) factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code (NTIS only).

Blocks 17.-19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contins classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

Abstract

This paper analyzes several theories of multi-agent collaboration. We have chosen three to study more closely than the others. While these contain what we believe to be the essential elements of a multi-agent collaboration, each approaches this idea somewhat differently. This is followed by a less-extensive review of several other theories, a number of which allow for contractual relationships. This analysis reveals not only relative strengths and weaknesses between the theories, but also illustrates the differences when agents collaborate and when they have a contractual association, the line between which is often blurred in the literature. Finally, we present two "real-world" domains, and discuss the advantages and disadvantages of these theories to each domain.

Table of Contents

Notation

The clauses within this document, while following the individual author's style and semantics, are written in a first order logic grammar; we assume the reader is familiar with the usual common connectives. The variables within these clauses are defined below, rather than in the body of this document. Because we will examine several theories, it is felt that a common lexicon in the front of this document will enhance readability and comprehension.

α, β	agents
Γ	action performed by the agent(s)
Δ	a condition that the agent(s) either believe, or have as a goal
ζ	context in which an agent performs some action Γ, or believes or has a goal Δ
τ	time
$\{x\}\|y$	action x with y holding initially. To enhance comprehension, the original notation [18] has been modified. The use of Latin instead of Greek letters here implies that either variable may be bound to either another variable or a predicate.

In addition to the notation, we also provide the reader with several expressions, reproduced from [18], that will aid in comprehension: if we have n agents α_n, and a sequence of events, e, then we specify the semantics of (AGT $\alpha_1 \ldots \alpha_n$ e) as $\alpha_1 \ldots \alpha_n$ are the only agents for e. We similarly define (HAPPENS a) and (HAPPENED a) as saying that some e describable by an action expression a will happen next, or has happened, respectively. Armed with these definitions, we can now define

(DOES $\alpha_1 \ldots \alpha_n$ a) _ (HAPPENS a) $\wedge$ (AGT $\alpha_1 \ldots \alpha_n$ a)

(DONE $\alpha_1 \ldots \alpha_n$ a) _ (HAPPENED a) $\wedge$ (AGT $\alpha_1 \ldots \alpha_n$ a)

(UNTIL x y) _ $\forall$z (HAPPENS z| ~{y}) $\supset$ $\exists$a (a $\leq$ z) $\wedge$ (HAPPENS a|{x})

where a $\leq$ z says that a is an initial subsequence of z.

Throughout this document the pronoun "his" is used for brevity only, and should be read as "his or hers".

I. INTRODUCTION

Applying artificial intelligence (AI) techniques to systems of greater complexity and sophistication places greater demands on both the knowledge content and processing speed of the executing software agents. In fact, these agents are increasingly being used in environments where knowledge is distributed throughout the agent pool, rather than being centralized. Distributed knowledge means that no one agent is capable of achieving the highest level goal(s) on his / her own; goal achievement becomes a team activity, where multiple agents collaborate towards the achievement of overarching goal(s). The advantages to this decentralization are increased processing speed and robustness; loss of an agent doesn't necessarily doom the activity, as the remaining agents could possibly "pick up the slack". On the other hand, the price for this increased robustness comes in the form of overhead related to team formation and collaboration, agent communication and team maintenance. What makes an agent want to start, or join a team, and once a team member, what obligations does it assume, either explicitly or implicitly? Answers to questions such as these fall within the domain of multi-agent collaboration (MAC), an area of active research within distributed artificial intelligence. Collaborating agents are either being used, or considered for use in many types of settings, running the gamut from everyday examples cited in [7] and the Internet [16], to multiple spacecraft formations in deep space [17].

This paper explores various MAC theories. In section II we focus on three influential contributions to the field: Joint Intentions, SharedPlans, and Planned Team Activity. The theory of Joint Intentions [4, 5] is the first serious attempt to formalize a collaborative theory. In addition, it explicitly includes several key concepts that one would intuitively associate with a multi-agent team, such as a shared responsibility towards other team members, and explicit communication requirements. The SharedPlan theory [9] has its origins in Pollack's theories [22]. Like Joint Intentions, its formulation includes concepts of mutual belief and the necessity of having all team members behave in a manner consistent with goal achievement, such as performing individual actions, and expecting others to do likewise. However, it differs from the former theory when it comes to communication and the requirement for a believable recipe for the team's action. Unlike the first two theories, Kinny's Planned Team Activity [14] explores the mechanics of team formation itself. This particular theory presents an interesting contrast to the first two: it lacks a notion of joint commitment, and it presupposes that all agents have complete knowledge of the goal and way(s) to achieve it prior to actual team formation.

We present other collaboration theories in section III. Their inclusion here is meant to illustrate that there is no unique definition of multi-agent collaboration about which a single theory can be developed. Two of these authors, Castelfranchi [1] and Matsubayashi and Tokoro [20], lean more towards contracting and delegating responsibility, respectively, rather than truly collaborating. Cavedon and Tidhar [2] lack an intention at the team level; their claim is that intention resides solely within individual agents. Huber and Durfee [11] relax the communication requirement to the extent that it's the responsibility of the other team members to infer an agent's intentions from his behavior, rather than the agent communicating his beliefs. Finally, Tidhar, *et al.* [29] allows for different levels of commitment, an option not allowed in either the Joint Intentions or SharedPlan theories. Each of these theories is discussed in more detail in the text, and they are contrasted with those in section II.

In the final analysis, the "goodness" of any MAC theory is domain-dependent. Section IV applies the theories in section II, primarily Joint Intentions and SharedPlans, to "real-world" domains, robotic soccer (Robocup [15]) and a NASA study of precision formation flying for long baseline interferometry by multiple spacecraft [17]. These are two domains that we find to be more appropriate to these MAC theories than the others. This paper then concludes in section V with a Summary and proposal for future work.

II. Three Theories of Multi-Agent Collaboration

We will compare and contrast three important models of multi-agent collaboration in this section. Each model will be seen to have its own unique strengths and weaknesses. These models discussed here form a baseline with which the other models in section III are compared.

Joint Intentions

The Joint Intentions model of Cohen and Levesque [4, 5, 6, 18] represents one of the first attempts to establish a formal theory of multi-agent collaboration, and due to its clarity and expression, is probably the best known of the rational agent theories. The basic premise of their theory rests on the idea of an *intention* - commitments (individual or joint) to act in a certain mental state [5, 18]. A *commitment*, in turn, represents a goal that persists over time [5]. One of the more interesting conclusions of this theory is that when an agent adopts, or comes to believe privately some notion that is not shared by the rest of the team, that agent has a responsibility to make this belief known to the rest of the team. In other words, by virtue of his having entered into a joint commitment with other agents, he has implicitly agreed to communicate his private beliefs if he believes the joint goal is either achieved, unachievable, or irrelevant. It was somewhat of a surprise to find that other MAC theories don't share the belief that agents must communicate with others. In particular, the theory of SharedPlans [9] suggests that whether or not an agent communicates is unpredictable, depending on "... other intentions and its beliefs."

The authors arrive at their theory of joint intentions by first defining the ideas behind individual commitment and intention. They are able to express these definitions in terms of a temporal logic augmented by propositional terms. An agent α has a persistent goal towards some goal Δ if the agent currently believes it not to be true, has a goal to eventually make it true, will retain such a goal until the agent either believes it true, believes it will never be true, or is irrelevant. This can be formally expressed[1] as

$$(\text{PGOAL } \alpha\ \Delta\ \zeta) \_ \ (\text{BEL } \alpha \sim\Delta) \wedge (\text{GOAL } \alpha \blacklozenge\Delta) \wedge (\text{KNOW } \alpha\ (\text{UNTIL } [(\text{BEL } \alpha\ \Delta) \vee (\text{BEL } \alpha \sim\Box\Delta) \vee (\text{BEL } \alpha \sim\zeta)]\ (\text{GOAL } \alpha \blacklozenge\Delta)))$$

The important point here is that an agent cannot just arbitrarily to drop a PGOAL; the *KNOW* clause contains the conditions under which this can be done. It will be seen below that, as in this individual case, the joint analog of a PGOAL is what serves to bind agents together, for once a joint commitment is adopted, it cannot just be arbitrarily abandoned.

Consistent with the concept of an agent behaving rationally is the notion that an agent act in a conscious, deliberate manner. Specifically, an agent can have a PGOAL towards some goal, and have that goal state realized by the agent behaving and performing actions in an arbitrary manner. This, however, is getting away from the spirit of describing agents that act deliberately and intentionally, and are cognitive of their actions. Hence, we assume our agents have a PGOAL to have done an action, and believe that he was doing it while he was doing it, or

$$(\text{INTEND } \alpha\ \Gamma\ \zeta) \_ \ (\text{PGOAL } \alpha\ (\text{DONE } \alpha\ \{\text{BEL } \alpha\ (\text{DOES } \alpha\ \Gamma)\}|\Gamma)\ \zeta)$$

[1] The following discussion can be found in [6]. It is included here for completeness.

We can then define the concept of a joint persistent goal, JPG, as[2]

$$(\text{JPG}\ \alpha\ \beta\ \Delta\ \zeta)\ \_\ \ (\text{MB}\ \alpha\ \beta\ \sim\Delta) \wedge (\text{MG}\ \alpha\ \beta\ \blacklozenge\Delta) \wedge$$
$$(\text{MK}\ \alpha\ \beta\ (\text{UNTIL}\ [(\text{MB}\ \alpha\ \beta\ \Delta) \vee (\text{MB}\ \alpha\ \beta\ \sim\Box\Delta) \vee (\text{MB}\ \alpha\ \beta\ \sim\zeta)]\ (\text{WMG}\ \alpha\ \beta\ \Delta)))$$

where having a mutual goal, MG, is defined by both agents believing that each agent independently desires to bring about some Δ:

$$(\text{MG}\ \alpha\ \beta\ \Delta\ \zeta)\ \ \_\ \ (\text{MB}\ \alpha\ \beta\ (\text{GOAL}\ \alpha\ \Delta\ \zeta) \wedge (\text{GOAL}\ \beta\ \Delta\ \zeta))$$

Additional replacements in the PGOAL clause that bring us to a JPG are mutual knowledge, MK (replacing KNOW), and weak mutual goal, WMG (replacing (GOALα $\blacklozenge\Delta$)).

$$(\text{WMG}\ \alpha\ \beta\ \Delta\ \zeta)\ \_\ \ (\text{MB}\ \alpha\ \beta\ (\text{WG}\ \alpha\ \beta\ \Delta\ \zeta) \wedge (\text{WG}\ \beta\ \alpha\ \Delta\ \zeta))$$

where

$$\begin{aligned}(\text{WG}\ \alpha\ \beta\ \Delta\ \zeta)\ \_\ \ & [\sim(\text{BEL}\ \alpha\ \Delta) \wedge (\text{GOAL}\ \alpha\ \blacklozenge\Delta)] \vee [(\text{BEL}\ \alpha\ \Delta) \wedge (\text{GOAL}\ \alpha\ \blacklozenge(\text{MB}\ \alpha\ \beta\ \Delta))] \vee \\ & [(\text{BEL}\ \alpha\ \sim\Box\Delta) \wedge (\text{GOAL}\ \alpha\ \blacklozenge(\text{MB}\ \alpha\ \beta\ \sim\Box\Delta))] \vee \\ & [(\text{BEL}\ \alpha\ \sim\zeta) \wedge (\text{GOAL}\ \alpha\ \blacklozenge(\text{MB}\ \alpha\ \beta\ \sim\zeta))]\end{aligned}$$

When two agents have a JPG they are mutually bound to bring about the goal, and if either agent develops private beliefs that the goal is either achieved, unachievable or irrelevant, he must make this private knowledge mutually believed.[3] Hence, the need for some form of communication is implicit in this model.[4] The form of this communication is domain dependent; in the case of a soccer team, communication could be either verbal or hand waving, using some prearranged signals. There exist other domains (see, *e.g.* [26, 27]) where, communication, while still vital to a team's success, has to be weighed against other factors, thereby revising the absolute requirement to communicate. The bottom line is that in all but the most unusual circumstances should an agent recognize his responsibility to make his private beliefs known regarding the team's commitment to the joint goal. Once an agent believes the goal is either achieved, unachievable or irrelevant, the JPG is dissolved; however the WMG remains, and it is through this goal that he is committed to making his beliefs known. That the JPG is dissolved when one of these conditions is valid can be seen by the following: a) if an agent believes Δ has been achieved, then (MB α β $\sim\Delta$) is false; b) if an agent believes Δ is unachievable, then (MG α β Δ ζ) and hence (MG α β $\blacklozenge\Delta$ ζ) is false; c) if an agent believes Δ is irrelevant, then the JPG fails for the same reasons as part b).

We would like to point out several minor differences between [18] and [6]. In the third clause of the JPG expression (Definition 4 in both papers), the former has "(UNTIL [(MB...)])", while the latter has "(MK α

---

[2]Several definitions in [6] are missing contextζ; we have modified them here.

[3]We do not address the question of*how* an agent adopts his private beliefs.

[4]Not all models share this point of view regarding communication. See, e.g. [11].

β (UNTIL [(MB...)])). The difference is more than just syntactic; without this MK expression, there is no requirement that they know that each has a WMG (to have one's private beliefs mutually believed). If I (as an agent) don't know that he knows he has a WMG, I might unnecessarily doubt his commitment given some unexplained behavior on his part, rather than relying on the MK that I will be informed of any changes in his beliefs. We believe, therefore, that the inclusion of this MK expression in the latter paper [6] corrects the earlier incorrect semantics of JPG. Another difference is in the MG clause ([18], Definition attempt, pg. 96) and ([6], Definition 4). The former contains the temporal qualifier ♦ (EVENTUALLY), where the latter does not. Putting no requirement on when an agent wants to (or has to) achieve some goal (as in (GOAL α Δ)), relaxes the semantics of the clause to the point that one could infer failure if Δ wasn't achieved now. Explicitly requiring a goal to be achieved eventually allows an agent to maintain a goal if not initially achieved. The remaining difference lies in the WG (weak goal) expressions, Definitions 3 and 5, respectively. In particular, the latter contains a final clause [(BEL α ~ζ) ∧ (GOAL α ♦(MB α β ~ζ))], which the former lacks. Without this clause, an agent that adopts a private belief that the joint action towards some goal is irrelevant will have no private goal of making his belief mutually known, thus allowing an agent to abandon his commitment to the joint goal with no further action towards the remaining team members.

SharedPlans

The SharedPlan model of collaboration [10] was developed to account for several deficiencies noted in Pollack's mental state of plans [21, 22], these being spelled out in [9]:

1) an action being performed by two or more agents could be decomposed into actions being performed by each individual agent. Grosz [7] provides a lucid discussion of why a joint action is not the sum of the individual plans, and hence why collaboration cannot be achieved by gluing separate plans together, but must be an integral part of the joint action from the beginning;

2) an agent was not required to establish a commitment towards the success of a collaborative partner's actions. Lacking such a commitment, an agent will be less likely to constrain its planning with regard to another agent's needs or activities. It incurs little or no obligation towards both sharing of global resources, and rendering assistance if the other agent should require it in the course of its actions in support of the joint goal;

3) mental states of agents while executing partial plans. Agents undertaking joint activities frequently do not have an entire roadmap for their activities throughout the course of their collaboration, or if they start with a fully-specified plan, an uncertain and dynamic world could obviate some fraction of that plan. Hence the need to maintain a collaboration in the face of incomplete plans.

Having a SharedPlan implies a joint mental state; this can be seen from its definition [19]: Both agents mutually believe that:

1) each is able to perform certain activities at the appropriate time;
2) there is a single recipe for performing the action;
3) each intends to perform his specific actions at the appropriate time;
4) each intends to perform their specific actions as a way of contributingo the overall action.

In their model, Grosz & Kraus address these issues with their specification of the key components of mental states of agents when they have a collaborative plan to do a group action:

1) mutual belief of a (partial) recipe;
2) individual intentions that the [joint] action be done;
3) individual intentions that collaborators succeed in doing the (identified) constituent subactions;
4) individual or collaborative plans for subactions

We note the following important differences between this model (which we label "Grosz"), and that of Cohen and Levesque ([6], hereafter known as "Cohen"):

1) These four components above lack the notion of a joint intention. This is a significant departure from Cohen , where the notion of joint intention is an integral part of their theory. In particular, #2 has agents individually intending that the joint action be done, while #3 has agents individually intending the success of their collaborators' actions (Grosz and Kraus introduce an *intention-that* operator to handle this, which will be discussed below).

2) There is no mechanism within these four components that makes an allowance for the departure of an agent's private beliefs from that of the group, and what his consequent responsibilities are, if any, towards the group. In Cohen, the agent has a commitment (weak goal) towards the other team members to make his private beliefs known.

3) Lastly, what becomes of the joint action should an agent no longer intend the joint action? In Cohen the joint intention vanishes; the fate of a joint action in Grosz remains unclear.

4) There is no requirement, either implicit or explicit, for agent communication in Grosz. In Cohen, an agent must communicate when he perceives that the joint goal is either achieved, unachievable, or irrelevant. This lack of a communication requirement in Grosz differs from her recent article [7], where she spells out very clearly the need for communication; in fact, she illustrates this need with several different domains.

As a way of describing "attitudes of intention"[5] four different intention operators: *intention to*, *intention that*, *potential intention to*, and *potential intention that* are introduced [8]. The first two are intentions that have been adopted by an agent. Potential intentions represent an agent's mental state when it is considering adopting an intention, but has yet to weigh the consequences of adopting that intention with others it presently holds. The idea here is that this motivates an agent to weigh different possible courses of action. The key point about *intentions to* and *intentions that* is that both commit an agent not to adopt conflicting intentions, and constrain replanning in case of failure [9].

An agent that has an *intention to* do something must believe two things:

1) it will be able to do the action at the appropriate time
2) it can *successfully* perform any action it intends to do.

This last condition is quite strong; the authors admit as much.[6] Any doubt in the agent's ability to perform

[5] [8], pg. 368.

[6] [9], pg. 11.

an action would preclude its even attempting the action.

In contrast with *intention to*, an *intention that* does not directly connote an action. Rather it implies that agents will behave in a manner consistent with a collaborative effort: they won't adopt intentions that conflict with the joint activity, and they will adopt intentions to communicate about the plan [9]. Communication requirements are derived from any *intention that*'s, as opposed to being "hard-wired" in Joint Intentions. Specifically, in the axiom of *Intending That* ([9], Axiom A5), if an agent has an *intention that* some group action succeed, and has adopted an *intention to* towards some subaction (as part of the larger group action) that it is, for some reason, unable to complete, it will adopt a *potential intention to* do any action that it believes will make the group action succeed as long as it still retains its original *intention that* the group action succeed. If that agent believes communicating its failure, beliefs, etc. will aid in successfully prosecuting the group action, then it will communicate.

A provision is made in this theory for contracting - having an agent performing some (or all) of the actions of another. An agent, for reasons of expediency or lack of knowledge, may wish to contract out some actions which he has previously agreed to undertake. According to their theory, if there is a recipe for performing some overall action that is known by all the participating agents, and if each agent is doing some portion of that overall recipe, then in a contracting situation, an agent may contract out some or all of his subactions to another, and the part of the recipe that contains the contractor's subactions is replace by the contracting action itself.

The basic difficulty with this is that contracting is not a commitment towards some joint goal, for the act of contracting does not entail that the contractee be in a "certain mental state..." [19], nor does it require the two agents to be in some joint mental state. It only requires that the contractee agree to accept some other agents' tasks, not to be in some mental state if, or when those tasks are being performed.

Earlier, it was stated that one of the components of mental states of agents when they have a collaborative plan to do a group action is a mutual belief of a (partial) recipe. This notion of a partial recipe (a/k/a *partial individual plan*) is important, for if an agent does not have a *complete* plan for performing some individual action, then it does not necessarily have an intention to either perform that action or contract it out. Possessing a complete plan, i.e. a complete recipe (called a *full individual plan*) implies the intention to perform the action(s). As opposed to a full individual plan, a partial individual plan (PIP) deals with situations in which agents have incomplete knowledge about how to perform some complex action. The only constraints on having a PIP to perform some action are that the agent "... just needs some idea of how to get a recipe" [9]; it doesn't even need a partial recipe for the action.

Communication remains an unresolved issue. Their approach is that if an agent decides to abandon a collaboration, it will adopt a potential intention to explain its decision, this (possible) communication being predicated on the agent's "... other intentions (*intentions that*) and its beliefs. Some people communicate in such situations; others do not" ([9], pg. 58). It is noteworthy to contrast how the SharedPlan formulation treats communication with that of the Joint Intention framework, where, in the latter case, communication is made an explicit requirement of a multi-agent team.

Planned Team Activity

The previous two models have implicitly assumed that when agents establish either a joint commitment or SharedPlan, they do so immediately and completely; there are no allowances for "expressions of interest": a stronger condition than "I'm not interested in forming a team", and a weaker condition than a full commitment. In this theory, Kinny [14] discusses two methods of team formation (details below) based on

a team leader and his communication with potential team members. An additional difference between this and the previous two theories is that plans to achieve some goal are supplied in advance, not generated by the agents, and that agents have complete knowledge of the full plans prior to their joining a team. An advantage to pre-specifying plans is that agent behavior is now bounded and predictable, and can respond readily and to a dynamic environment. On the other hand, agent behavior is more brittle, as replanning is compromised in an unpredictable environment. In addition, there is greater responsibility placed on the agent designer in that their success is heavily dependent on how well the plans have been specified and mirror anticipated conditions.

The semantics of teams's beliefs, goals and intentions are different from those in Cohen's Joint Intentions [6]. Specifically, the joint attitudes of a team are expressed in terms of the joint attitudes of its members which reduce to single agent attitudes, rather than by modal operators that express attitudes held by everyone in the team.[7] A team has a joint intention towards a plan if:[8] (a) every member has the joint intention towards the plan; (b) every member believes that the joint intention is held by the team; and (c) every member believes that all members executing their respective individual plans results in the team executing the joint plan.

We note the similarity with Grosz's definition of plans for collaborative action ([7], fig. 8). The additional feature that Grosz captures is the intention that collaborators succeed. This is critical, for without this condition agents could behave in such a way as to inadvertently interfere with the progress of other team members, such as exclusive use of what should really be a shared resource.

The process of team formation begins with an agent wishing to achieve some goal, but realizing that he is unable to do so by himself. He then goes about assembling a team, with himself installed as the team leader. He communicates with other potential participants by announcing the joint goal, joint plan, and the individual roles to be assumed by each participant. A team member is capable of adopting a joint goal, a plan, and a role within a plan if and only if:[9] (a) team has the necessary skills; (b) team member does not already believe the formula that needs to be adopted as a joint goal; (c) preconditions of the plan are already believed by the team; (d) joint goal is compatible w/ the current goals of the team member; and (e) joint plan and role plan are compatible with the current intentions of the team member, where a plan is considered compatible with existing intentions if it does not conflict with those conditions required to maintain those intentions.

Two versions of team formation are considered here: 1) commit-and-cancel, and 2) agree-and-execute.[10] While there are differences between these two strategies, what they do share is the fact that the joint goal, joint plan, and individual roles are known prior to an agent's commitment and subsequent team formation.

Commit-and-Cancel: The team leader sends a request to each participant to "commit" to the joint goal, joint plan, and role. The role is implicit in the instantiated joint goal. If the team

---

[7] [14]. Specifically, Definition 10 defines joint beliefs, goals and intentions of single agents in terms of individual attitudes. With these definitions, they define joint beliefs, goals and intentions of teams as the conjunction of the joint attitudes of the agents and the mutual belief that particular attitude (belief, goal or intention) is held by the rest of the team.

[8] [14], p. 253.

[9] [14], p. 248.

[10] The following description of these two algorithms can be found in [14], pp. 250-251. Milind Tambe recognized that these two strategies are very similar to eager and lazy protocols, respectively.

leader receives a "committed" reply from all the participants then the team has been formed, with the joint goal and joint intention being simultaneously established, and execution begins immediately. If any one of the participants does not commit, or doesn't reply within the permitted time, the team leader must send an explicit "cancel" message to each agent that has committed. Upon receipt of this message, the participating agent deletes its intention and stops executing the plan.

Agree-and-Execute: The team leader sends an "agree" request to all participants, and if all reply affirmatively, sends an explicit request to all of them to execute the plan. It is at this point that the joint plan, joint goal, and individual roles are adopted. Unlike Commit-and-Cancel, no explicit message needs to be sent if one of the participants does not agree to participate, as the other agents have made no commitment to the goal prior to the "execute" message from the team leader.

If a member is unable to execute its part of the plan, then it has the responsibility to make other team members aware of its failure. This isn't surprising; what is surprising is that the theory requires this agent to coordinate the response to its failure. What this suggests is that the team leader's role is completed once the team is formed and has begun execution. This is okay if there is only one failure; multiple failures require a central coordination mechanism that should logically remain within the purview of the team leader. In any event, this can be a retry, or another plan which can satisfy the desired goal. If a plan fails, the team may retry the plan with a different role assignment before admitting failure.

Team members are able to consider more attractive opportunities as they arise in the course of a dynamic world. As a consequence of this reconsideration, a team member may abandon the joint intention.; it must, as before, communicate this to other team members. This stands in marked contrast to Cohen's theory of joint intentions that allows an agent to abandon a joint goal only if it comes to privately believe that the goal is either achieved, unachievable or irrelevant. In addition, in Cohen's theory, communication is not enough; the agent must communicate to others with the weak goal of having its beliefs mutually believed, at which time the joint goal is abandoned. In the present theory, an agent's responsibility is discharged following this communication, and the mutually-held beliefs held by the other are not necessarily abandoned.

In summary, this model would work well in a predictable environment with "well-qualified" agents, as plans are pre-enumerated and the strategies of team formation require that agents be aware of both these plans, as well as their own capabilities. In a dynamic and unpredictable domain, team formation via this theory would be a more risky proposition, as plans are subject to change once adopted, and the more opportunistic team members can opt out of a joint commitment by merely announcing their intention to do so.

III. Other Models of Multi-Agent Collaboration

Many social agency models were reviewed for this paper. The three that were presented in the previous section were selected based on their originality, contrast with each other, and acceptance by the multi-agent community. We gain a broader understanding of the field, however, if we broaden our definition of multi-agent collaboration to include theories that, in addition to contracting and delegating, present opposing points of view regarding joint intentions and communication.

Castelfranchi [1] argues that a deeper notion of commitment includes an agent's mental states, and

relationships to other agents; these mental states bind one or more agents to each other. He distinguishes three types of commitments: *internal* (or I-Commitments), *social* (or S-Commitments), and *collective* (or C-Commitments). I-Commitments are what an individual agent intends towards an action, S-Commitments are relational; their semantics require two or more agents as expressed by (S-COMM $\alpha\ \beta\ \gamma\ \delta$), where agent α is committed to agent β to perform some action γ, or (S-COMM $\alpha\ \beta\ \gamma\ \delta$)$\supset$ (GOAL β (DOES $\alpha\ \gamma$)). Castelfranchi notes that this a form of "goal adoption": by virtue of the right-hand side of this equation, β has γ as a goal. Agent δ is a "witness", before whom α has an implied commitment, and is not considered further in [1]. Just as individual agents can hold individual intentions through I-commitments, a group of agents can hold a C-Commitment, which expresses the group's internal commitment towards some act. Castelfranchi imbues I-Commitments with a semantics similar to that of [4, 18]: I-Commitments are persistent, and will be abandoned only when an agent comes to believe a goal is achieved, unachievable, or irrelevant.

By adopting an S-commitment towards β, α pledges (has an I-Commitment) to perform some specific act, β comes to expect this, can require that α does it and can complain when he doesn't. In other words, certain specific rights are conferred on β; in acknowledgment of these rights, α, as part of its S-Commitment, won't oppose these rights [1]. While one might be tempted to relate these Commitments to the intention operators *intention-to* and *intention-that* in [8, 9], their semantics precludes this kind of association. In the present work, Commitments speak of cooperation, not collaboration. Agents that collaborate seek to achieve a common goal while in a "shared mental state", a quality lacking in a cooperative agreement. As such, when an agent holds an S-Commitment towards another, this more closely resembles a contractual arrangement, one agent agreeing to perform some act for another (for whatever reason). Consequently, the agent whose work is being done for him: a) adopts the goal that the agreeable agent will perform the specific act, not he; and b) will not interfere, nor do anything that would prevent the "doing" agent from achieving his goal. Both agents are not actively working towards, or collaborating on achieving the goal, rather there is a cooperative agreement implied by the adoption of the S-Commitment by both agents; this cooperation is spelled out in the semantics of joint adoption of the S-Commitment by both agents. In fact, because one agent is acting in the service of another, this is really a contract protocol. The *intention* operators referenced above are cast in a collaborative framework. For example the *intention-that* clause assumes that "...agents [will] adopt intentions to communicate about the plan and its execution" [9, p. 36], the implication being that both agents take an active role towards the achievement of a joint goal. Grosz [7] lays out the semantics of *intending-that* very succinctly when she relates this operator to joint activity; there is no such activity in an S-Commitment; intention in this operator does not imply a joint activity. Lastly, both this theory and that of Joint Intentions [6] do not allow a joint goal to be dropped merely by abandoning a private goal; obligations among the remaining team members are incurred (through an S-Commitment) that must be satisfied.

Castelfranchi raises an apparent paradox in his paper: two scientific competitors have the same goal (to find a vaccine for AIDS), but because they are competitors no one would (rationally) claim they form a team. Yet (according to Castelfranchi) they satisfy the three JPG clauses (see pg. 2), hence would be considered a team in the Joint Intentions framework [1, p. 46]. If we look more closely at what it means to be form a JPG, we will see that there is no contradiction at all.[11] A JPG is a conjunction of three clauses[12], one of which is a mutual goals clause:

$$(\text{MG}\ \alpha\ \beta\ \Delta\ \zeta)\ \_\ (\text{MB}\ \alpha\ \beta\ (\text{GOAL}\ \alpha\ \Delta\ \zeta) \wedge (\text{GOAL}\ \beta\ \Delta\ \zeta))$$

[11] Milind Tambe (private communication) first pointed this out to the author, and provided useful insight into the overall problem.

[12] See section on Cohen and Levesque for definitions.

where Δ is the goal to be achieved. Now the two researchers have the same apparent goal, namely to find a vaccine. Yet the goal is not the same; each one wants to achieve it individually, not jointly. So there is no shared belief state in achieving the same goal, hence no joint intention [6]. Further, since the goals are dissimilar, in the above equation for MG, $\Delta \rightarrow (\Delta_1 \wedge \Delta_2)$, where α and β have goals Δ_1 and Δ_2, respectively. Making this substitution, the right-hand side becomes:

$$(\text{MB } \alpha\ \beta\ (\text{GOAL } \alpha\ (\Delta_1 \wedge \Delta_2)\ \zeta) \wedge (\text{GOAL } \beta\ (\Delta_1 \wedge \Delta_2)\ \zeta))$$

Since a goal, say Δ_1 is not accessible in β's world, it is not believable, hence β cannot have it as a goal.[13] A similar argument holds for Δ_2 and α. So neither agent cannot have both Δ_1 and Δ_2 as goals, there is no MG, hence no JPG. Searle [24] also argues that joint intentions are not built from the conjunction of individual intentions.

Cavedon and Tidhar [2] present a theory quite different from those previously discussed. They use a belief-desire-intention (BDI) architecture to describe their agents. They are then able to relate these attributes at the agent level to those at the team level. This specification can be performed in either a top-down (or *reductionist*) manner towards the team members, or a bottom-up (or *holistic*) manner from the members to the team [23]. The primary advantage of the former specification is that intentions could be assigned at the highest, or team, level, thereby obviating the need to enumerate the lowest level intentions of the individual agents. They would "get their marching orders" from individual BDI libraries.

A bottom-up specification is thought to be a more realistic representation of the natural world, and is referred to as "emergent behavior". As opposed to a top-down system, an emergent system produces, from the rudimentary behaviors at the agent level, more complex behaviors at the team level. In particular, in an emergent system, a team-level intention is not imposed on the system, rather it arises from the interaction of the individual agents [2]. A consequence of this is that team members are not aware of any team-level intention.

Lastly, there is the issue of when to drop intentions. According to this theory, an agent drops his intentions when the goal is "... believed to hold." [2, p. 11]

This theory lacks many of the features found in the others, namely, the mechanics of team formation, intentions of team members towards a joint goal, the notion of jointness towards both goals and intentions, communication, and a more rigorous theory of how, and under what conditions an agent can abandon his private goals, his responsibilities towards the other team members, and the state of the team following this goal loss.

Recall that in the Joint Intentions framework, when a JPG exists between two agents and one comes to privately believe that the joint goal is either achieved, unachievable or irrelevant, he has a weak goal to make this fact mutually believed. It is his responsibility to communicate with the other team members. In Huber [11], the authors relieve the agent of the communication task, and instead place the responsibility for determining the continuing viability of the joint goal on the team members. An agent still may communicate if so inclined, it's now longer a requirement. In lieu of communication, the individual members may gather whatever information they can from the environment and interpret it in a subjective manner.

[13] See definitions in [5, p. 7]

This kind of paradigm is obviously quite different from the previous theories presented. There are several practical issues that could offset the potential benefit of reducing agent overhead by eliminating communication requirements. For example, what happens if a team member believes that the agent that just dropped his commitment to the joint intention had an absolutely critical skill that, without which, the joint goal is unachievable. He would then seriously reconsider my continued belief in the joint intention, whereas another agent might have completely different beliefs, and *he might be right.* It is this agent's *incorrect* perception that led to his private belief of a joint goal doomed to failure, when in actuality his information was faulty. Nonetheless, he might be responsible for the team dropping its joint intention. An excellent example of what could go wrong when an agent observes, draws inferences, and acts (incorrectly) instead of communicating directly is given as a convoy example in [5]. This theory would be valuable in a domain where communication is either unreliable, or where it must be abandoned altogether (see, *e.g.* [26, 27]).

Matsubayashi [20] explores the relationship between two agents, one of whom wishes to delegate his responsibilities to another. This delegation arises because in the development of an individual agent's own actions and goals, there is bound to be overlap with another agent's if both are part of a team with the same overarching goal(s). The idea is that delegation of these overlapping parts can reduce the overall execution cost. To avoid one agent being delegated too many responsibilities, a protocol, or "social law" is setup. Its purpose is to balance the numbers of delegated / received actions for each agent. Adherence to this social law provides a guarantee that agents won't be "taken advantage of", and avoids (or at least reduces) the need for conflict resolution.

The delegation phase is not unilateral assignment; it is really a negotiation phase where agent α proposes that agent β receive a subgoal of α's for him (β) to execute. Both agents then seek to minimize cost / maximize payoff through this negotiation phase.

However, life being what it is, one can expect to encounter "self-centered" agents, or those that do not abide by this social law. The problem arises when one of these sociopathic agents has some other quality (or ability) that would make collaboration with it beneficial to the team. Negotiation with one of these self-centered agents is not as problematic with this protocol, because each time there is a conflict, the probability of this self-centered agent abiding by the law is reduced. Since negotiation will only occur where a positive payoff (and adherence to the law) is expected, there is an increasingly less chance that a self-centered agent will be negotiated with, but that chance doesn't vanish altogether. Agents may become more or less self-centered, depending on the goals and their available options.

This theory is one of a contractual arrangement, rather than a collaboration. It quantifies the benefits to the team given the delegation of tasks that could be "more easily" performed by another team member. Yet in a collaboration, we would expect to see some notion of a joint goal, responsibility of one team member to another and to the team itself, and how to deal with the consequences of an agent's inability to perform the action he has agreed to undertake.

Tidhar [29], in a manner similar to Kinny [14], explores how a team is actually formed. He presents several aspects of team formation, the underlying ideas of which stand in contrast to others theories we have so far reviewed. The exception to this is [14], where the same protocols related to team formation are presented, and which will not be repeated here.

One of the central ideas of their thesis is that plans are built-in within each agent; this is in deference to operating in a real-time domain where time is at a premium. While minimizing deliberation time within each agent, the obvious price to be paid for this deterministic behavior is a rather brittle plan, although one can envision domains where this would be satisfactory; the authors give an example of transformer

maintenance in a power grid [29]. In any event, agents operate in a plan space where each agent has a library of plans that represent the goal decomposition into achievable subgoals. The intersection of these individual libraries represents the team plan library. A team is considered capable of adopting this plan when a given set of conditions are satisfied [29, p. 7]. From these conditions, what's interesting to note is that, in contrast with both the Joint Intentions and SharedPlans formulations, there is no notion of: a) a joint intention towards a goal, or b) a shared responsibility towards the successful outcome of the plan. Lacking these semantics, this theory doesn't possess the richness of either of the two previous ones. Developing a notion of jointness is central to the idea of a collaboration, as this is expressed through some good examples in both [5] and [7]. Underscoring this lack of jointness is the fact that the authors allow for what they call "different grades of commitment." It is this author's personal belief that modulating one's commitment to a joint goal is inconsistent with being a team member. One is either committed to a goal, or one is not. A team composed of agents that are less than fully committed to a goal is likely to fail when confronted with an adverse circumstance.

IV. Applications

While the theories are all well and good, and have their own internal consistency, in the final analysis what matters is how well-behaved they are when applied to real-world problems. The ultimate benchmark is, of course, how humans would "do it"; after all, since these software agents are deemed rational, and endowed with several human-like attributes, it is reasonable to assume that their behavior (which is proportional to their sophistication) asymptotically approaches that of a human. We will look at two different domains: Robocup soccer, and NASA proposals to deploy multiple spacecraft formations. These domains were chosen primarily for personal reasons: the former domain served as an entrée to the study of MACs, while the latter is closely related to the author's work at the Jet Propulsion Laboratory. These domains actually are interesting because they represent real-world applications of MACs; they are not idealized examples designed to illustrate any one theory. Additionally, there are more than two agents composing these teams. A two-team scenario doesn't capture the richness or challenges posed by a multi-agent team. For example, if one of the agents on a two-agent team (for whatever reason) abandons the joint goal, regardless of any remaining commitments, the team is effectively dissolved. Not necessarily so with a larger team; the distributed nature of a team's activities could allow another member, or members, to assume the (sub) goals of the departing agent. All aspects of team formation, communication and intention are still held by the remaining members.

MACs could obviously be applied to more domains than just the two mentioned here. A common application is a cooking domain that involves just two agents [9]. Recently, Tambe [26, 27, 28] analyzed air combat scenarios with both fighters and helicopters based on the Joint Intentions framework as presented in [5]. In particular, he demonstrated the importance of teamwork to team tracking - inferring another team's goal and intentions by correlating observations of another team's behavior with an internal model of how the observing agent(s) believe it should behave. Let us now examine our domains.

Robocup

First introduced in 1995 [15], this is a soccer competition that will be played[14] between two teams of soccer agents on a simulated soccer field. This domain bears some similarities to Tambe's air combat scenarios [26, 27, 28] where there are two teams, both mutual antagonists. The soccer domain, both real as

[14]The First Robot World Cup (RoboCup-97) in Nagoya, Japan, August, 1997.

well as robotic, poses additional challenges because of the speed of the action. Its dynamic nature limits replanning, and while winning the match is obviously the overarching goal, the number of agents actually involved with any particular scoring opportunity is some subset of the entire team. Because different agents are involved with various on-field activities at different times, this means that agents of different abilities and beliefs are constantly forming and breaking little "subteams" very quickly. The requirement for communication and the ability to quickly infer a teammate's intentions is very high.[15] As currently proposed, both the Planned Team Model and SharedPlans would not be as natural a fit in this domain as Joint Intentions. The Planned Team Model [14] assumes that individual agent roles are known prior to an agent's commitment and subsequent team formation. As previously mentioned, soccer's dynamic nature involves the rapid restructuring of subteams, precluding *a priori* knowledge of a particular agent's role within the subteam. The Planned Team Model would have to relax this requirement of "built-in" plans to be effective in this domain. The SharedPlans formulation [9] lacks an explicit communication requirement; communications are a consequence of agents' intentions, rather than being made a requirement of their actions. In soccer, the need for on-field communication is very high, given the dynamic nature of the game; any theory likely to be applied to this (or any other rapidly-changing) team sport must include an explicit requirement that agents communicate. We would propose the following modification to SharedPlans for this domain: extend the communications allowance in *intention that* to a requirement in *intention to*. In this manner, when an agent instantiates an intention to a personal action, he notifies the others. Knowledge of this agent's intentions and actions guides adoption of *intention that*'s by the others. They will not interfere with that agent's actions, assist if necessary, and can replan if that agent notifies the others that he has either finished, or is unable to achieve his goal.

Applying the Joint Intentions theory [4, 5, 6, 18] to our soccer domain addresses the concerns left by the other theories. One area that would have to be augmented, however, deals with an agent's persistence in trying to achieve a goal. Recall that when a team has a joint persistent goal (JPG) to do something, each agent has a goal to eventually make that "something" true. Since it's not reasonable to expect an agent to pursue a goal indefinitely, in spite of the agent's beliefs, there must be some heuristics that augment the theory that tell an agent when to abandon a goal. According to this theory, he still has a responsibility to make his beliefs mutually known, so that the others don't pursue a goal that an agent has opted out of.

As currently written, the Joint Intentions theory is a two-agent theory. In particular, two agents appear in the clauses which define mutual goal, weak mutual goal, and weak goal. In some soccer activities, such as a wall pass, these definitions are adequate. Many defensive alignments, however, require more than two players. We must modify these definitions to accommodate more than two players. In our original description of this framework we had only two agents on a team, α and β. We'll modify our notation slightly to accommodate additional agents. A team will be composed of n agents, α, which we'll denote as α_n. We then modify our earlier definitions as follows:

$$(\text{JPG } \alpha_n \Delta \zeta) \quad \_ \quad (\text{MB } \alpha_n \sim\Delta) \wedge (\text{MG } \alpha_n \blacklozenge\Delta) \wedge (\text{MK } \alpha_n (\text{UNTIL } [(\text{MB } \alpha_n \Delta) \vee (\text{MB } \alpha_n \sim\Box\Delta) \vee (\text{MB } \alpha_n \sim\zeta)] (\text{WMG } \alpha_n \Delta)))$$

$$(\text{MG } \alpha_n \Delta \zeta) \quad \_ \quad (\text{MB } \alpha_n [\bigwedge^{n}_{i=1} (\text{GOAL } \alpha_i \Delta \zeta)])$$

Weak mutual goal (WMG) was originally defined in terms of two agents: the conjunction of two clauses,

[15] [25]. The context in which Stewart explicitly mentions communication deals with covering and marking (defense); communication is obviously just as vital while on offense.

each .of which specifies that one agent has a "weak goal" to achieve some goal relative to the other agent [6]. With *n* agents, we will have a total of n(n-1) clauses, analogously denoting that each agent has a "weak goal" relative to every other agent. We write this as

$$(\text{WMG } \alpha_n \Delta \zeta) \_ \quad (\text{MB } \alpha_n [\bigwedge^{n}_{i=1} (\text{WG } \alpha_i \alpha_{j|j=1,n,j\neq i} \Delta \zeta))])$$

In a similar fashion, we write for weak goal,

$$\begin{aligned}(\text{WG } \alpha_n \Delta \zeta) \quad \_ \quad & [\sim(\text{BEL } \alpha_i \Delta) \wedge (\text{GOAL } \alpha_i \blacklozenge \Delta)] \vee [(\text{BEL } \alpha_i \Delta) \wedge (\text{GOAL } \alpha_i \blacklozenge (\text{MB } \alpha_n \Delta))] \vee \\ & [(\text{BEL } \alpha_i \sim\Box\Delta) \wedge (\text{GOAL } \alpha_i \blacklozenge (\text{MB } \alpha_n \sim\Box\Delta))] \vee \\ & [(\text{BEL } \alpha_i \sim\zeta) \wedge (\text{GOAL } \alpha_i \blacklozenge (\text{MB } \alpha_n \sim\zeta))]\end{aligned}$$

where the subscript *i* denotes the i^{th} agent within the *n*-agent team. It will be noted that these general definitions reduce to the original definitions when n=2.

Multiple Spacecraft Formations

The days of interplanetary spacecraft in the Galileo and Cassini class costing billions of dollars and consuming thousands of person-years to both develop and operate are over. NASA has shifted its priorities in favor of smaller and cheaper spacecraft, albeit with no compromise of science data return. Accomplishment of these objectives requires the use of technologies that have never before been used in a deep space spacecraft. The current NASA program that will use these technologies is the New Millennium program (NMP), in which AI will play a prominent role.[16] Presently undergoing a feasibility study is an NMP optical interferometer project known as Deep Space-3 (DS-3). This is a proposal to fly a small fleet of three identical spacecraft in a tightly-controlled formation, and ultimately validate similar missions comprised of larger assemblages of spacecraft, or constellations.[17] Such a proposal is not new; multiple spacecraft interferometers were proposed in the early 1980's [3], but the technology was not there to support such a mission. With the advent of sophisticated hardware and software, these missions are now considered feasible.

In DS-3, each spacecraft will have identical capabilities for command and control, with one of the three spacecraft designated as the "master". Communication to and from Earth, as well as with the other formation members will be through this master. In addition, initializing and maintaining the formation geometry and orientation changes of the formation will take place with respect to the master [17]. Operationally, the master will merely rotate, while the others both translate and rotate.

Conceptually, this formation architecture can be thought of as "quasi-centralized"; one of the formation members is the designated master, yet either of the other members would have the capability to assume the master role should the current master suffer some unpredictable problem that would render it unable to act as either the communication, or command and control hub. Each formation member would be responsible, however, for maintaining its own attitude as per instructions from the master.

---

[16]The first New Millennium flight, DS-1, will have an onboard "Remote Agent" consisting of a RAPS-based executive, an HSTS-based planner/scheduler, and a Livingstone-based mode identification and recovery algorithm.

[17]Such a mission has been proposed. Known as MUSIC (multiple spacecraft interferometer constellation), its objective will be to establish a long baseline for optical interferometry with a constellation of 16 spacecraft.

The ability to assume the master's role by any of the members at any time requires that the software be architecturally identical in all spacecraft. Homogeneous software not only guarantees identical capabilities in all spacecraft, but also considerably simplifies spacecraft functionality for critical functions, such as fault protection [17]. In addition, because the software is identical in all spacecraft, only one working "version" need be developed, rather than separate software development efforts for each spacecraft, resulting in a considerable cost savings. The spacecraft must be bound in a collaborative effort, otherwise the spacecraft have no obligation to either each other or the overall mission. A collaboration implies responsibilities among the formation's members, such as communication when an anomaly is perceived or the goal threatened, and not acting in a competitive manner with other members for shared resources, such as communication time or data storage on the master.

Of the theories we have reviewed which one is best for this domain? No one theory has all the elements that would be suffice for this application. The Joint Intentions framework would have to be augmented with some domain-dependent heuristic that would tell the master when to stop communicating with a "sick" comrade, or when one of the formation members should stop trying to achieve an attitude that would jeopardize its consumables (power or fuel). Goal abandonment would probably be based upon some previously-enumerated set of conditions. In any event, whenever a goal (such as reorientation) has been achieved, or is deemed impossible, the other members are informed, not just the master. The reason is the following: let's assume that one of the formation members (say x_1) hasn't achieved its prescribed attitude, the error is just beyond the acceptability boundary. It adopts the belief that its goal is unachievable, and consequently seeks to achieve its weak goal of making this condition mutually believed. This is not *necessarily* a reason to drop the JPG. One of x_1's associates, x_2, realizes that if he reorients himself just off his nominal attitude, he can compensate for the error induced by x_1's failure, and so maintain the JPG. In the case of three total spacecraft in the formation, and two could equally-well compensate for the third, the decision as to who would do the actual reorienting could be based on some heuristic hierarchy, the highest one being, for example, that the one who has the greatest fuel reserves would do the honors.

In the case of the SharedPlans formulation, we previously discussed the lack of an explicit communication requirement. It is the author's opinion that this is a critical capability that must be a part of any collaboration in this particular domain. Space is too dynamic and unpredictable an environment to leave autonomous agents functioning with their own notions what others intend to do. Explicitly communicating ones intentions and beliefs minimizes the guesswork in an environment where incorrect decisions could spell substantial financial loss.

V. Summary

Many of the future AI applications will be in domains where several agents are working together towards a common goal. We have explored some of the MAC theories, the ideas behind how agents form teams, how the team is maintained along with each agent's obligation as a team member, and what happens when a team member wants to "bail out." We have explored the high level ideas of three theories: Joint Intentions, SharedPlans, and Planned Team Activity. While these theories have some features in common, they also differ in rather significant ways, most notably with communication and team dissolution. In the interest of exploring the breadth of this field, we also briefly examined some other ideas that address multi-agent collaboration. We have observed that several of the theories interleave the notions of contracting, cooperating and collaborating. When agents contract, one is operating in the services of another. Independent of whether responsibility is delegated ("you do it") or assumed ("I'll do it for you"), there is no "shared mental state" between the agents towards the joint achievement of a goal, event though they might have a mutual interest in seeing it achieved. Cooperation is a superset of contracting, because in

addition to contracting, it also includes a passive acceptance of an agent's goals, *e.g.*"I don't want to help, but I won't interfere either." Again, what separates cooperation from collaboration is a shared mental state. Finally, we took two sample domains, and discussed how these theories might fit into such domains, and any shortcomings regarding their applicability. Almost all tasks in the real world involve more than one agent. This reflects the complexity of the world, limited knowledge and/or capabilities of the agents, and the realization that, more often than not, the most expeditious way to achieve a goal is to through teamwork, and the distribution of the various subtasks among interested and capable agents.

Among the theories we have studied, Joint Intentions comes the closest to what this author would consider the best theory for our two example domains. It has intuitive appeal, and presents teams as robust and dedicated, comprised of agents with a sense of responsibility both to themselves and others. This is not to say that it's perfect, however. We can borrow some ideas from other theories we have studied that could augment the Joint Intentions (JI) framework. For example, there are domains where this hard-wired communication requirement in JI could be a liability [28]. Inferring an agent's intentions from its actions by others would take the place of communicating. In a real-time environment, the deliberation time is limited, for an agent needs time to act prior to the next "cycle." Given this limitation, it would be advantageous for each agent to have a plan library it could invoke in a highly-constrained situation. Finally, with different knowledge and ability levels, it may be advantageous for an agent to contract out those actions for which he has assumed responsibility to others. The sum of the contracting overhead plus the cost of the first agent doing the task himself could be more than offset by another agent that enters into a contractual agreement with the first. Utilizing his specialized knowledge and / or resources, he could perform the task more efficiently, thereby reducing the overall cost of accomplishing the goal.

What AI has taught us so far is that there is no one way to do anything; a theory that's advantageous in one domain might be second-rate in another. It's up to the designer's discretion to apply the theory that is best in his or her domain. It's no different with MACs. We have attempted to paint a broad picture of the work that has been done, as well as the individual shortcomings that have to be addressed. In the final analysis, any one of these theories could provide an adequate basis on which to further develop the theory for a specific application. By striving to develop software agents that collaborate, we are really modeling our own behavior, which, as we know, is a very difficult task.

Acknowledgments

This Directed Research project has been quite an educational experience. It has been my good fortune to explore multi-agent collaborations with Milind Tambe. His knowledge, insight and persistence have been invaluable guides along the way.

Appreciation is expressed to the following people: Karen Lochbaum and Hector Levesque for their private communications, and Ken Lau at JPL for providing information on formation flying.

Finally, I wish to thank my family, Diane and Eric, for their understanding and patience with my highly-constrained schedule. The benefits will remain long after the work is done. This has truly been a multi-agent collaboration; one could not ask for a better team.

VI. REFERENCES

[1] Castelfranchi, C. "Commitments: From Individual Intentions to Groups and Organizations", In *Proc. 1st Intl. Conf. on Multi-Agent Systems*, AAAI Press, San Francisco, (1995), 41-48.

[2] Cavedon, L. and G. Tidhar. "A Logical Framework for Multi-Agent Systems and Joint Attitudes", *Australian Artificial Intelligence Institute Tech. Note 66*, Aug., 1995. This paper will also be published in *Proc. Dist. AI Workshop of the 8th Australian Jnt. Conf. on AI*, Canberra.

[3] Colavita, M., C. Chu, E. Mettler, M. Milman, D. Royer, S. Shakian, and J. West. "Multiple Spacecraft Interferometer Constellation (MUSIC): A New Concept for Astrophysical Imaging and Characterizing Planets Around Nearby Stars, *Jet Propulsion Laboratory*, JPL D-13369, February, 1996.

[4] Cohen, P. and H. Levesque. "Intention is Choice with Commitment", *Artificial Intelligence* **42**(3), (1990), 213-261.

[5] Cohen, P. and H. Levesque. "Teamwork", *Noûs* **25**(4), (1991), 487-512.

[6] Cohen, P. and H. Levesque. "Confirmations and Joint Action", *IJCAI-91*, (1991), 951-957.

[7] Grosz, B. "Collaborative Systems", *AI Magazine* **17**(2), (1995), 67-85.

[8] Grosz, B. and S. Kraus. "Collaborative Plans for Group Activities", *Proc. 1993 Intl. Joint Conf. Artificial Intelligence*, 367-373.

[9] Grosz, B. and S. Kraus. "Collaborative Plans for Complex Group Action", *Harvard University Report* #TR-20-95, 1995.

[10] Grosz, B. and C. Sidner. "Plans for Discourse", in P. Cohen, Morgan, J. and Pollack, M., eds., Intentions in Communication, Bradford Books, MIT Press, 1990.

[11] Huber, M. and E. Durfee. "On Acting Together: Without Communication", *Proc. AAAI Spring Symposium on Reasoning About Mental States*, 1995.

[12] Jennings, N. "Controlling Cooperative Problem Solving in Industrial Multi-Agent Systems Using Joint Intentions", *Artificial Intelligence* **75**, (1995), 195-240.

[13] Jennings, N. "Commitments and Conventions: The Foundation of Coordination in Multi-Agent Systems", *The Knowledge Engineering Review* **8**(3), (1993), 223-250.

[14] Kinny, D., M. Ljungberg, A. Rao, E. Sonenberg, G. Tidhar, and E. Werner. "Planned Team Activity", in *4th European Workshop on Modeling Autonomous Agents in a Multi-Agent World (MAAMAW)*, 1992.

[15] Kitano, H., M. Asada, Y. Kuniyoshi, I. Noda and E. Osawa. "Robocup: The Robot World Cup Initiative", in *Proc. IJCAI-95 Workshop on Entertainment and AI/Alife*

[16] Kouzes, R., J. Myers and W. Wulf. "Collaboratories: Doing Science on the Internet", *Computer*

29(8), (1996), 40-46.

[17] Lau, K. (ed.). "Coordinated Flying",*Jet Propulsion Laboratory*, March, 1996.

[18] Levesque, H., P. Cohen and J. Nunes. "On Acting Together",*AAAI-90*, 94-99.

[19] Lochbaum, K., B. Grosz and C. Sidner. "Models of Plans to Support Communication: An Initial Report", *Proc. AAAI-90*, 485-490.

[20] Matsubayashi, K. and M. Tokoro. "A Collaboration Mechanism on Positive Interactions in Multi-Agent Environments", *Proc. Intl. Joint Comm. On AI (IJCAI)*, 1993, 346-351.

[21] Pollack, M. "A Model of Plan Inference that Distinguishes Between the Beliefs of Actors and Observers", *Proc. 24th Ann. Mtg. Assoc. for Computational Linguistics*, (1986), 207-214.

[22] Pollack, M. "Plans as Complex Mental Attitudes", in P. Cohen, Morgan, J. and Pollack, M., eds., Intentions in Communication, Bradford Books, MIT Press, 1990.

[23] Rao, A. and M. Georgeff. "BDI Agents: From Theory to Practice", In *Proc. 1st Intl. Conf. on Multi-Agent Systems*, AAAI Press, San Francisco, 1995, 312-319.

[24] Searle, J.R. "Collective Intentionality". In P.R. Cohen, J. Morgan, and M.E. Pollack, eds., *Intentions in Communication*, M.I.T. Press, Cambridge, MA., 1990.

[25] Stewart, P. Way to Play Soccer, Carlton Books, Ltd., Prima Publishing, Rocklin, CA., 1995.

[26] Tambe, M. "Tracking Dynamic Team Activity", *Proc. Natl. Conf. Artificial Intelligence (AAAI)*, 1996.

[27] Tambe, M. "Teamwork in Real-World, Dynamic Environments", *Second Intl. Conf. on Multi-Agent Systems (ICMAS)* (to be published), 1996.

[28] Tambe, M. "Executing Team Plans in Dynamic, Multi-Agent Domains", in *Proc. AAAI Fall Symposium on Plan Execution: Problems and Issues*(1996).

[29] Tidhar, G., A. Rao, M. Ljungberg, D. Kinny, and E. Sonenberg. "Skills and Capabilities in Real-Time Team Formation",*Australian Artificial Intelligence Institute*, Tech. Note 27, July, 1992.